PRECONDITIONING THE AUGMENTED LAGRANGIAN METHOD FOR INSTATIONARY MEAN FIELD GAMES WITH DIFFUSION*

ROMAN ANDREEV†

Abstract. We discuss the application of the augmented Lagrangian method to the convex optimization problem of instationary variational mean field games with diffusion. The problem is first discretized with space-time tensor product piecewise polynomial bases. This leads to a sequence of linear problems posed on the space-time cylinder that are second order in the temporal variable and fourth order in the spatial variable. To solve these large linear problems with the preconditioned conjugate gradients method we propose a preconditioner that is based on a temporal transformation coupled with a spatial multigrid. This preconditioner is thus based on standard components and is particularly suitable for parallel computation. It is conditionally parameter-robust in the sense that the condition number of the preconditioned system is low for sufficiently fine temporal discretizations. Numerical examples illustrate the method.

Key words. mean field games, ADMM, ALG2, augmented Lagrangian, space-time, discretization, preconditioning, multigrid, B-splines

AMS subject classifications. 35K45, 49J20, 49M29, 65F10, 65M12, 65M55, 65M60, 65N22, 65Y05, 91A10, 91A13

1. Introduction. Mean field games and related models describe a wide range of social phenomena such as crowd motion, opinion dynamics, vaccination rates, stability of marriage, percolation of innovation, etc., and, moreover, appear as the equations to be solved in each time-step of the so-called JKO time-stepping scheme for gradient flows. In the instationary stochastic case introduced in [23], mean field games is a coupled system of a transport-diffusion equation for a density (of crowd, opinion, etc.) with a nonlinear equation for the value function running in the opposite temporal direction. Existing numerical methods for mean field games are based for instance on finite volumes [1, 2, 3], the dynamic programming principle [13, 16], or on convex duality [9]; further references can be found in [10]. In this work we reconsider the convex duality formulation of [14] and the ALG2 splitting method of [19, 9], a.k.a. ADMM, in the presence of diffusion. The idea to use ALG2 in this context seems to go back to [8], but other proximal type methods could be used instead [26, 12]. The algorithm generates a sequence of linear partial differential equations posed on the space-time cylinder that are second order in time and fourth order in space. We discretize them with space-time finite elements. To solve the resulting large linear algebraic problems with an iterative method such as conjugate gradients, we develop a preconditioner using a) the principle of operator preconditioning, b) a temporal transformation that block-diagonalizes the preconditioner, and c) a multigrid for the spatial blocks. This preconditioner is thus based on standard components and is particularly suitable for parallel computation. Our space-time discretization corresponds in a sense to continuous Galerkin time-stepping for the value function, which implies a loss of stability if the temporal resolution is too coarse [5]. For this reason, the proposed preconditioner is robust in the relevant parameters *provided* the temporal resolution is fine enough, see Figure 1. A slight modification of the preconditioner partly alleviates this restriction, see Figure 2.

The paper is structured as follows. In Section 2 we introduce the mean field games model, its convex formulation and the ALG2 method. In Section 3 we describe the discrete version of ALG2. In Section 4 we comment on preconditioning of the space-time linear problems. The

*May 1, 2017

†Univ Paris Diderot, Sorbonne Paris Cité, LJLL (UMR 7598 CNRS), F-75205 Paris, France (roman.andreev@upmc.fr)

numerical experiments in Sections 5.2–5.4 illustrate these concepts for 1d and 2d examples. The final Section 5.5 extends the model to include intermediate instantaneous costs.

Supporting Matlab code is available from https://github.com/numpde/alg2mfg/.

2. Mean field games and ALG2 with diffusion.

2.1. Mean field games. The spatial domain $D \subset \mathbb{R}^d$ is assumed to be a cuboid but the main ideas are generic. Let $T > 0$, set $J := (0, T)$. Consider the system of partial differential equations

$$\mathrm{KFP}[\rho, \phi] := \partial_t \rho - \nu^2 \Delta \rho + \mathrm{div}(\rho \nabla H(t, x, \nabla \phi)) = 0, \tag{1a}$$

$$\mathrm{HJB}[\rho, \phi] := \partial_t \phi + \nu^2 \Delta \phi + H(t, x, \nabla \phi) = A'(t, x, \rho), \tag{1b}$$

$$\text{s.t.} \qquad \rho(0) = \rho_0 \quad \text{and} \quad \phi(T) = -\Gamma'(x, \rho(T)), \tag{1c}$$

for $(t, x) \in J \times D$. We will refer to this system as the mean field games (MFGs) equations. See Section 2.2 for an interpretation of the equations. We omit the dependence on (t, x) in the notation where convenient (in particular, the right-hand-side of (1b) actually means $A'(t, x, \rho(t, x))$, etc.). Here, A and Γ are convex real valued functions of the third variable $\rho \geq 0$ that evaluate to $+\infty$ for negative ρ, and the indicated derivatives are with respect to this variable. The essential assumption on the Hamiltonian H is convexity with respect to the third variable. The unknowns are the density ρ and the cost ϕ, both space-time dependent real valued functions. Periodic boundary conditions are often assumed in the literature but here we will be interested in no-flow boundary conditions on the density ρ. Based on (1a), these are implemented by requiring

$$\nabla H(t, x, \nabla \phi) \cdot \mathbf{n} = 0 \quad \text{and} \quad \nabla \rho \cdot \mathbf{n} = 0 \quad \text{on} \quad \partial D \tag{2}$$

where $\mathbf{n}$ is the outward normal to the spatial boundary. In particular, the total mass $\int_D \rho$ is conserved in time, and we suppose that $\rho_0 \geq 0$ non-trivially. We consider only radially symmetric coercive Hamiltonians (in the third variable); hence the first condition of (2) amounts to

$$\nabla \phi \cdot \mathbf{n} = 0 \quad \text{on} \quad \partial D. \tag{3}$$

Let $L(t, x, \mathbf{v})$ be the Lagrangian obtained from the Hamiltonian H as the dual conjugate with respect to the third variable. By convexity of H in that variable, H is also the dual conjugate of L (and the formal optimality condition for the supremum is $\mathbf{p} = \nabla L(\mathbf{v})$):

$$H(\mathbf{p}) = L^*(\mathbf{p}) := \sup_{\mathbf{v} \in \mathbb{R}^d} \{\mathbf{p} \cdot \mathbf{v} - L(\mathbf{v})\}. \tag{4}$$

The principal feature of the MFGs equations is that the Kolmogorov–Fokker–Planck (KFP) equation evolves forward in time with an explicit initial condition at $t = 0$ and the Hamilton–Jacobi–Bellman (HJB) equation evolves backward in time with a possibly implicit initial condition at $t = T$. The mathematical interpretation of the KFP is in the weak sense and that of the HJB is in the viscosity sense, but we will mostly proceed in a formally (in particular assuming the regularity (11)–(12) below).

The main innovation of this work with respect to the numerical method based on finite elements proposed in [9] is the presence of the positive diffusion coefficient

$$\nu > 0. \tag{5}$$

We will therefore pay particular attention to robustness with respect to ν. The first consequence of (5) is a minor adaptation in the formulation due to the ν term in (14). The second consequence is the necessity of choosing the finite element spaces appropriately, e.g. in our case we require H^2-conformity of the spatial discretization of ϕ. The third consequence is the appearance of a linear partial differential operator $\mathscr{A}$ (defining successive approximations of ϕ in the ALG2 iteration) that is second order in time and fourth order in space. This means that a large linear algebraic system of equations, typically highly ill-conditioned, has to be solved in each outer iteration. The main contribution of this paper is a preconditioner that breaks that linear problem down into a sequence of independent spatial problems, each amenable to multigrid. The preconditioner is based on the principle of operator preconditioning [22, 25], i.e. it is obtained as the discretization of an operator $\mathscr{C}$ that is similar to $\mathscr{A}$ on the continuous level. We first motivate the functional framework in Sections 2.2–2.4, and then define the operator $\mathscr{C}$ in Section 2.5.1.

2.2. Variational formulation. As already observed in [23], the MFGs equations (1) arise as the first order optimality conditions of a constrained optimization problem which we recall here. This perspective provides some useful intuition for the quantities appearing in (1), see also the example in Section 5.2.

Consider the transport equation (slightly overloading the notation from (1a))

$$\mathrm{KFP}[\rho,\mathbf{v}] := \partial_t\rho - \nu^2\Delta\rho + \mathrm{div}(\rho\mathbf{v}) = 0. \tag{6}$$

One may think of the space-time dependent function ρ as the state variable and of the space-time dependent vector field $\mathbf{v}$ as the control variable. To fix ρ and $\mathbf{v}$ we introduce the functional (in Section 5.5 we also consider a variant with multiple Γ-like terms)

$$J_1(\rho,\mathbf{v}) := \int_{J\times D}\{L(\mathbf{v})\rho + A(\rho)\} + \int_D \Gamma(\rho(T)), \tag{7}$$

and the constrained optimization problem

$$\inf_{\rho,\mathbf{v}} J_1(\rho,\mathbf{v}) \quad \text{s.t.} \quad \mathrm{KFP}[\rho,\mathbf{v}] = 0. \tag{8}$$

For $\rho \le 0$, by convention, $L(\mathbf{v})\rho = 0$ if $(\rho,\mathbf{v}) = 0$ and $L(\mathbf{v})\rho = +\infty$ else. To formally characterize the minimizers we look for the stationary points of the space-time Lagrangian $J_1(\rho,\mathbf{v}) + \langle\phi, \mathrm{KFP}[\rho,\mathbf{v}]\rangle$, where the space-time scalar function ϕ is the Lagrange multiplier for the KFP constraint. The derivative of the Lagrangian with respect to $\mathbf{v}$ gives the relation $\nabla L(\mathbf{v}) = \nabla\phi$, at least where $\rho \neq 0$. Using the optimality condition for the supremum in (4), this implies the representation $H(\nabla\phi) = \mathbf{v}\cdot\nabla\phi - L(\mathbf{v})$ and hence the feedback strategy $\mathbf{v} = \nabla H(\nabla\phi)$. Employing the latter in the derivative of the space-time Lagrangian with respect to ρ gives the HJB equation (1b) with its terminal condition from (1c).

In the next subsection, following [8, 23, 14, 9], we use convex duality theory to obtain a formally equivalent formulation (15) of (8). First, some notation is in order.

2.3. The predual problem. We will work with the function spaces

$$\mathsf{H} := L_2(D), \quad \mathsf{V} := H^1_{\mathrm{Neu}}(D), \quad \mathsf{W} := H^2(D)\cap\mathsf{V}. \tag{9}$$

We write $\|\cdot\|_{J\times D}$ / $\|\cdot\|_D$ to indicate the $L_2(J\times D)$ / $L_2(D)$ norm. On V, which incorporates homogeneous Neumann boundary conditions, we use the norm given by

$$\|\chi\|_{\mathsf{V}}^2 := \|\chi\|_D^2 + \nu^2\|\nabla\chi\|_D^2, \quad \chi\in\mathsf{V}. \tag{10}$$

Introducing the Bochner–Sobolev space

$$\mathsf{X} := L_2((0,T);\mathsf{W}) \cap H^1((0,T);\mathsf{H}), \tag{11}$$

we suppose that the Lagrange multiplier ϕ from the previous subsection satisfies

$$\phi \in \mathsf{X}. \tag{12}$$

We abbreviate $L_2 := L_2((0,T)\times D)$. Recall the isometry $L_2 \cong L_2((0,T);L_2(D))$. We identify $\mathsf{H} \cong \mathsf{H}'$ via the Riesz isomorphism, obtaining the Gelfand triple $\mathsf{V} \hookrightarrow \mathsf{H} \cong \mathsf{H}' \hookrightarrow \mathsf{V}'$, and in particular the embedding $\mathsf{V} \hookrightarrow \mathsf{V}'$. Set

$$\mathsf{Y} := L_2 \times L_2^d \times \mathsf{V}. \tag{13}$$

This choice of spaces and norms will be important in Section 2.5.1 in the derivation of the preconditioner $\mathscr{C}$. We identify again the L_2 spaces with their dual. Elements of Y will be denoted by $\sigma = (a,\mathbf{b},c)$ or $\lambda = (\rho,\mathbf{m},e)$, and those of Y' by $\lambda' = (\rho,\mathbf{m},e')$. We define the linear operator

$$\Lambda : \mathsf{X} \to \mathsf{Y}, \quad \phi \mapsto (\partial_t\phi + \nu^2\Delta\phi, \nabla\phi, -\phi(T)). \tag{14}$$

This definition follows [9] except that here we admit $\nu > 0$.

The operator Λ is injective. Indeed, multiplying $\partial_t\phi + \nu^2\Delta\phi = 0$ by ϕ, integrating over $J\times D$, integrating by parts on Δ, and using $(\partial_t\phi,\phi)_{\mathsf{H}} = \frac{1}{2}\partial_t\|\phi\|_{\mathsf{H}}^2$ gives $\|\phi(t)\|_{\mathsf{H}} \le \|\phi(T)\|_{\mathsf{H}}$. In particular, $\Lambda\phi = 0$ implies $\phi = 0$, which is the raison d'etre for the third component of Λ. The key observation that was made and exploited in [23, 14, 9] is that the optimization problem

$$\boxed{\operatorname*{argmin}_{(\phi,\sigma)\in\mathsf{X}\times\mathsf{Y}} \{\mathscr{F}(\phi) + \mathscr{G}(\sigma)\} \quad \text{s.t.} \quad \sigma = \Lambda\phi} \tag{15}$$

with

$$\mathscr{F}(\phi) := \int_D \phi(0)\rho_0, \quad \mathscr{G}(\sigma) := \int_{J\times D} A^\star(a + H(\mathbf{b})) + \int_D \Gamma^\star(c) \tag{16}$$

is a reformulation of the MFGs equations (1), because it is in duality with the constrained optimization problem (8). Here, $(\cdot)^\star$ denotes the convex conjugate, i.e.

$$A^\star(z) = \sup_\rho\{z\rho - A(\rho)\}, \quad \Gamma^\star(c) = \sup_\rho\{c\rho - \Gamma(\rho)\}. \tag{17}$$

By assumptions on A and Γ, the functions $A^\star$ and $\Gamma^\star$ are "flat" on negative arguments.

For convenience of the reader, and to highlight some technical points, we explain how (15) and (8) are connected in the remainder of this subsection. First, by the general theory of duality [18], the (typical) convex dual problem of (15) reads

$$\inf_{\lambda'\in Y'} \{\mathscr{F}^\star(-\Lambda'\lambda') + \mathscr{G}^\star(\lambda')\}, \tag{18}$$

where $\Lambda' : Y' \to X'$ is the adjoint (see Theorems 1–2 below for the relation between the minimizers of (15) and (18)). We write $[L+A]$ for the function $(\rho,m) \mapsto L(\mathbf{m}/\rho)\rho + A(\rho)$. Using $H(b) = \sup_{\mathbf{m}}\{b\cdot\mathbf{m}/\rho - L(\mathbf{m}/\rho)\}$ one finds $A^\star(a+H(\mathbf{b})) = [L+A]^\star(a,\mathbf{b})$. Under standard

conditions [18, Prop. 4.1], the function $[L+A]$ coincides with its second dual. The same holds for Γ. This leads to (cf. (7))

$$\mathscr{G}^\star(\lambda') \stackrel{\text{def}}{=} \sup_\sigma\{\langle\sigma,\lambda'\rangle - \mathscr{G}(\sigma)\} = \int_{J\times D}\{L(\mathbf{m}/\rho)\rho + A(\rho)\} + \int_D \Gamma(e'). \tag{19}$$

Concerning $\mathscr{F}^\star$, we first observe (using integration by parts in time)

$$\langle -\Lambda\phi,\lambda'\rangle = \langle\phi,\partial_t\rho - \nu^2\Delta\rho\rangle - \langle\phi,\rho\rangle|_{t=0}^{t=T} + \langle\phi,\operatorname{div}\mathbf{m}\rangle + \langle\phi(T),e'\rangle - \mathrm{BT}, \tag{20}$$

where the spatial boundary terms are collected in

$$\mathrm{BT} := \int_{J\times\partial D}\big(\nu^2\rho\nabla\phi - (\nu^2\nabla\rho - \mathbf{m})\phi\big)\cdot\mathbf{n}. \tag{21}$$

Invoking the spatial homogeneous Neumann boundary conditions (3) we find that

$$\mathscr{F}^\star(-\Lambda'\lambda') = \sup_\phi\{\langle -\Lambda\phi,\lambda'\rangle - \mathscr{F}(\phi)\} \text{ is finite (and equals zero)} \tag{22}$$

if and only if the equations

$$\partial_t\rho - \nu^2\Delta\rho + \operatorname{div}\mathbf{m} = 0, \quad \rho(0) = \rho_0, \quad \nu^2\nabla\rho\cdot\mathbf{n} = \mathbf{m}\cdot\mathbf{n} \quad\text{and}\quad \rho(T) = e' \tag{23}$$

are satisfied in the weak sense. In view of (19) and (22), writing $\mathbf{v} := \mathbf{m}/\rho$ we find that (18) is just the constrained optimization problem (8). We note here two consequences of (23) for the numerical method in Section 3 below:

- The spatial Neumann boundary conditions (2) imply $\mathbf{m}\cdot\mathbf{n} = 0$ on ∂D, which we also use for the discrete flux $\mathbf{m}_h$.
- From the discrete triple $\lambda_h = (\rho_h, \mathbf{m}_h, e_h)$, an approximation of the velocity field in (8) may be obtained as $\mathbf{v}_h := \mathbf{m}_h/\rho_h$.

2.4. On existence of minimizers. The following two classical results [18, Remark 4.2] concern the optimization problems (15)–(18).

THEOREM 1. *Suppose $\mathscr{F}$ and $\mathscr{G}$ are convex. Suppose there exists ϕ_0 such that $\mathscr{F}(\phi_0)$ and $\mathscr{G}(\Lambda\phi_0)$ are finite and $\mathscr{G}$ is continuous at $\Lambda\phi_0$. Then* $\inf(15) = \inf(18)$ *and* (18) *has a minimizer.*

THEOREM 2. (ϕ,λ') *solves* (15)–(18) *and* $\inf(15) = \inf(18)$ *iff* $(-\Lambda'\lambda',\lambda') \in \partial\mathscr{F}(\phi)\times\partial\mathscr{G}(\Lambda\phi)$.

Typically, $A^\star$ and $\Gamma^\star$ will be convex nondecreasing functions, and together with convexity of H those conditions suffice to show convexity of F and G directly. Formally, the two inclusions from Theorem 2 correspond to the KFP (1a) and the HJB (1b) equations, respectively.

The built-in regularity assumption $\phi\in\mathsf{X}$ in (15) implies that $\phi(0)\in\mathsf{V}$, so that $F(\phi)$ is well-defined even for $\rho_0\in\mathsf{V}'$. Let $\sigma = (a,b,c) = \Lambda\phi$. Then $a\in L_2$, $b\in\mathsf{V}^d$ and $c\in\mathsf{V}$, and certain growth conditions on $A^\star$, $\Gamma^\star$ and H ensure that $G(\sigma)$ is well-defined. Let us focus on $d=2$ spatial dimensions. Then we may assume quadratic growth for $A^\star$ and $\Gamma^\star$ and arbitrary polynomial growth for H in the last variable. For an arbitrary $\sigma\in\mathsf{Y}$ not in the range of Λ we may extend the definition of G by $+\infty$ but this would preclude the usage of Theorem 1. Alternatively, we could restrict Y, say $\mathsf{Y} = L_2\times L_2((0,T);H^1(D))\times\mathsf{V}$, but we chose not follow this road in this paper in order to simplify the numerics in Step B of ALG2 below. Another possibility is to assume convexity and at most quadratic growth for $A^\star$ and $A^\star\circ H$ ensuring

continuity of G on Y so that Theorem 1 could be used but this would rule out the desirable situation of the standard quadratic Hamiltonian H together with a quadratic cost A in (7). In any case Theorem 1 does not address existence for the primal problem (15), and seeing (18)–(15) as the primal–dual pair is problematic because $F^\star$ is discontinuous wherever it is finite (unless again, we modify Y suitably).

Existence and stability in space-time Lebesgue spaces were obtained using Theorem 1 in [14], and by operating directly on the MFGs equations for example in [24, Theorem 2.7] and [3, Theorem 3.1]. The methods of [3] do not seem to apply in our case due to the lack of a discrete maximum principle.

Another possible approach to existence of minimizers in (15) is to verify coercivity of the functional under suitable assumptions on the data (which would also play a role in establishing Γ-convergence of numerical solutions [9, Section 3.1]). This, however, does not seem possible because the term $\int_D \Gamma^\star(-\phi(T))$ does not provide control on the spatial derivatives of $\phi(T)$, which would be necessary to control the norm of ϕ in X (see e.g. [6, Theorem 4.1] for that kind of statement). We believe the mesh-dependent convergence rate of ALG2, see Section 5.4, is a manifestation of this fact.

2.5. ALG2 formulation. In order to solve the optimization problem (15), in the augmented Lagrangian method one looks for saddle points $(\phi,\sigma,\lambda)\in \mathsf{X}\times\mathsf{Y}\times\mathsf{Y}$ of the *augmented Lagrangian*

$$L_r(\phi,\sigma,\lambda) = \mathscr{F}(\phi)+\mathscr{G}(\sigma)+(\Lambda\phi-\sigma,\lambda)_{\mathsf{Y}}+\tfrac{r}{2}\|\Lambda\phi-\sigma\|^2_{\mathsf{Y}}, \tag{24}$$

where $r\geq 0$; these are characterized by

$$L_r(\phi,\sigma,\cdot)\leq L_r(\phi,\sigma,\lambda)\leq L_r(\cdot,\cdot,\lambda). \tag{25}$$

It is elementary to check that the Lagrangians L_r and L_0 have the same saddle points. Moreover, any saddle point furnishes a solution to (15) and (18). With the notation of the previous subsection it would have been natural to have the duality pairing $\langle\Lambda\phi-\sigma,\lambda'\rangle$ with $\lambda'\in\mathsf{Y}'$ instead of the Y scalar product but below it is more convenient to work with the Riesz representative $\lambda\in\mathsf{Y}$ given by $(\lambda,\cdot)_{\mathsf{Y}}=\lambda'$. With the L_2 identification already made in Section 2.3, the first two components of λ and λ' coincide.

Let $r>0$. Consider the function $h_r(\lambda):=\inf_{\phi,\sigma}L_r(\phi,\sigma,\lambda)$. Then λ in the saddle point characterization (25) maximizes this function. At any point $\bar\lambda$, its gradient direction is $(\Lambda\bar\phi-\bar\sigma)$ where $(\bar\phi,\bar\sigma):=\operatorname{argmin}L_r(\cdot,\cdot,\bar\lambda)$. This suggests a steepest ascent algorithm for finding the optimal λ. This is the algorithm "ALG1" in [19, Section 3.1]. The subsequent algorithm "ALG2" [19, Section 3.2] is a modification where the minimization of ϕ and σ is decoupled. It proceeds by iterating the following three steps.

A. Minimize L_r with respect to the first component by solving the elliptic problem

$$\text{Find}\quad \phi^{(k+1)}\in\mathsf{X}\quad\text{s.t.}\quad \langle\tfrac{\partial L_r}{\partial\phi}(\phi^{(k+1)},\sigma^{(k)},\lambda^{(k)}),\tilde\phi\rangle=0\quad\forall\tilde\phi\in\mathsf{X}. \tag{26}$$

B. Proximal step:

$$\sigma^{(k+1)}:=\operatorname*{argmin}_{\sigma\in\mathsf{Y}}\{G(\sigma)+\tfrac{r}{2}\|\bar\sigma^{(k+1)}-\sigma\|^2_{\mathsf{Y}}\}\quad\text{for}\quad\bar\sigma^{(k+1)}:=\Lambda\phi^{(k+1)}+\tfrac{1}{r}\lambda^{(k)}. \tag{27}$$

C. Multiplier update:

$$\lambda^{(k+1)}:=\lambda^{(k)}+r(\Lambda\phi^{(k+1)}-\sigma^{(k+1)}). \tag{28}$$

The algorithm is also known as ADMM or Douglas–Rachford, and it was shown in [17] to be a proximal point algorithm. A recent overview of convergence results may be found in [20]. We now comment on each of those steps separately, and use this occasion to introduce the operator that will be the basis for the preconditioner below.

2.5.1. Step A. Since $\mathscr{F}$ is linear, (26) amounts to the linear variational problem

$$\text{Find} \quad \phi^{(k+1)} \in \mathsf{X}: \quad (\Lambda\phi^{(k+1)}, \Lambda\tilde{\phi})_{\mathsf{Y}} = (\sigma^{(k)} - \tfrac{1}{r}\lambda^{(k)}, \Lambda\tilde{\phi})_{\mathsf{Y}} - \tfrac{1}{r}\mathscr{F}(\tilde{\phi}) \quad \forall \tilde{\phi} \in \mathsf{X}. \tag{29}$$

The only term involving the unknown $\phi^{(k+1)}$ is (we suppress the iteration superscript)

$$(\Lambda\phi, \Lambda\tilde{\phi})_{\mathsf{Y}} = \langle(\partial_t + \nu^2\Delta)\phi, (\partial_t + \nu^2\Delta)\tilde{\phi}\rangle + \langle\nabla\phi, \nabla\tilde{\phi}\rangle + (\phi(T), \tilde{\phi}(T))_{\mathsf{V}}. \tag{30}$$

Expanding the first term on the right-hand side we obtain

$$\langle(\partial_t + \nu^2\Delta)\phi, \ldots\rangle = \langle\partial_t\phi, \partial_t\tilde{\phi}\rangle + \nu^4\langle\Delta\phi, \Delta\tilde{\phi}\rangle - \nu^2\langle\nabla\phi(t), \nabla\tilde{\phi}(t)\rangle|_{t=0}^{t=T}, \tag{31}$$

having used integration by parts in time and space on the term $\langle\partial_t\phi, \Delta\tilde{\phi}\rangle$. The boundary term $\nu^2\int_{J\times\partial D}(\phi\nabla\tilde{\phi} - \tilde{\phi}\nabla\phi)\cdot\mathbf{n}$ disappears for any combination of homogeneous/periodic Dirichlet/Neumann spatial boundary conditions. The negative ν^2 term cancels by the definition of the norm on V, so we are left with

$$\|\Lambda\phi\|_{\mathsf{Y}}^2 = \|\partial_t\phi\|_{J\times D}^2 + \nu^4\|\Delta\phi\|_{J\times D}^2 + \|\nabla\phi\|_{J\times D}^2 + \|\phi(T)\|_D^2 + \nu^2\|\nabla\phi(0)\|_D^2. \tag{32}$$

Consider the operator $\mathscr{A} := \Lambda'\Lambda : \mathsf{X} \to \mathsf{X}'$, where the adjoint is with respect to the Y scalar product; hence $\langle\mathscr{A}\phi, \tilde{\phi}\rangle = (32)$. Below we will iteratively invert a discretized version of $\mathscr{A}$, and will therefore require a good preconditioner. To that end we define the symmetric operator $\mathscr{C} : \mathsf{X} \to \mathsf{X}'$ by omitting the last term in (32), cf. Section 4.1, i.e.

$$\langle\mathscr{C}\phi, \phi\rangle := \|\partial_t\phi\|_{J\times D}^2 + \|\phi(T)\|_D^2 + \|\nabla\phi\|_{J\times D}^2 + \nu^4\|\Delta\phi\|_{J\times D}^2. \tag{33}$$

The operator $\mathscr{C}$ is equivalent to $\mathscr{A}$ by an argument similar to that of [4, Section 2.5]. Specifically, take the eigenfunctions φ_n normalized to $\|\varphi_n\|_D = 1$ and the corresponding eigenvalues $0 = \mu_0 < \mu_1 \le \ldots$ of the operator $-\Delta$ with the Neumann boundary conditions (3). Consider the expansion $\phi = \sum_n \theta_n \otimes \varphi_n$. The first inequality in the equivalence (33) $\lesssim$ (32) $\lesssim$ (33) is obvious. For the second, we need to estimate $\nu^2\|\nabla\phi(0)\|_D^2 = \sum_n \nu^2\mu_n|\theta_n(0)|^2$ in terms of $(33) = \sum_n\{\|\theta_n'\|_J^2 + |\theta_n(T)|^2 + (\mu_n + \nu^4\mu_n^2)\|\theta_n\|_J^2\}$. It suffices to find a constant $K > 0$ such that $K^{-1}|f(0)|^2 \le \alpha^{-1}\|f'\|_J^2 + \alpha^{-1}|f(T)|^2 + \alpha\|f\|_J^2$ uniformly in $\alpha > 0$ and $f \in H^1(J)$, where α and f are placeholders for $\nu^2\mu_n$ and θ_n. We consider two cases:

- If $\alpha \le 1$, we can use e.g. $|f(0)| \le \sqrt{T}\|f'\|_J + |f(T)|$ to infer a suitable K.
- If $\alpha \ge 1$, we assume w.l.o.g. $f(0) = 1$. Minimizing $\alpha^{-1}\|f'\|_J^2 + \alpha\|f\|_J^2$ over f leads to the boundary value problem $-\alpha^{-1}f'' + \alpha f = 0$ with $f(0) = 1$ and $f'(T) = 0$. The exact solution (a catenary) yields $\alpha^{-1}\|f'\|_J^2 + \alpha\|f\|_J^2 = \tanh(\alpha T)$. This suggests $K := \tanh T$.

In both cases we obtain a constant K that depends only on the time horizon T. Consequently, the equivalence (32) $\sim$ (33) is robust in the diffusion coefficient ν.

The variational problem (29) is convenient for finite element discretization as is. In strong form (29)–(31) amounts to a PDE of the form $(-\partial_t^2 + \nu^4\Delta^2 - \Delta)\phi^{(k+1)} = \text{RHS}$, which shows that the problem is of second order in time and of fourth order in space.

2.5.2. Step B. Recall the notation $\sigma = (a, \mathbf{b}, c)$ and $\lambda = (\rho, \mathbf{m}, e)$. Step B consists of two decoupled proximal subproblems, the first one being

$$(a^{(k+1)}, \mathbf{b}^{(k+1)}) := \operatorname*{argmin}_{(a,\mathbf{b})\in L_2\times L_2^d} \left\{ \int_D A^\star(a + H(\mathbf{b})) + \tfrac{r}{2} \|(\bar{a}^{(k+1)}, \bar{\mathbf{b}}^{(k+1)}) - (a, \mathbf{b})\|^2_{L_2\times L_2^d} \right\} \tag{34}$$

with the "priors" $\bar{a}^{(k+1)} := (\partial_t + \nu^2\Delta)\phi^{(k+1)} + \frac{1}{r}\rho^{(k)}$ and $\bar{\mathbf{b}}^{(k+1)} := \nabla\phi^{(k+1)} + \frac{1}{r}\mathbf{m}^{(k)}$, and the second one being

$$c^{(k+1)} := \operatorname*{argmin}_{c\in V} \left\{ \int_D \Gamma^\star(c) + \tfrac{r}{2} \|\bar{c}^{(k+1)} - c\|^2_{\mathsf{V}} \right\} \tag{35}$$

with the "prior" $\bar{c}^{(k+1)} := -\phi^{(k+1)}(T) + \frac{1}{r}e^{(k)}$. The first subproblem is as in [9, Section 4.2] but the second is not due to the non-L_2 norm, which is a consequence of the choice (13) of Y.

2.5.3. Step C. Step C is a straightforward update.

3. Discretization.

3.1. Discrete spaces. We discretize the quantities (ϕ, σ, λ) in (24) using tensor products of piecewise polynomial functions on an interval. We use the following notation, leaving the underlying mesh to be specified separately:

P1. Continuous piecewise affine functions (hat functions);
P2. Continuous piecewise quadratic functions;
D0. Piecewise constant functions (top functions);
D1. Piecewise polynomials of degree one with no inter-element continuity;
B2. Piecewise polynomials of degree two with C^1 continuity.

For simplicity, we assume here that the spatial dimension is $d = 2$. We write $\mathrm{P2}\otimes\mathrm{B2}^2$ for the function space spanned by the products of P2 functions in time with B2 functions in each spatial dimension, etc. We will look for a discrete saddle point $(\phi_h, \sigma_h, \lambda_h)$ as follows:

$$\phi_h \in \mathsf{X}_h := \mathrm{P2}\otimes\mathrm{B2}^2 \tag{36a}$$

$$a_h, \rho_h \in \mathsf{A}_h := \mathrm{D1}\otimes\mathrm{D0}^2 \tag{36b}$$

$$\mathbf{b}_h, \mathbf{m}_h \in \mathsf{B}_h := \mathrm{D0}\otimes[(\mathrm{P1}\otimes\mathrm{D0})\times(\mathrm{D0}\otimes\mathrm{P1})] \tag{36c}$$

$$c_h, e_h \in \mathsf{C}_h := \mathrm{B2}^2. \tag{36d}$$

Furthermore, we set $\mathsf{Y}_h := \mathsf{A}_h \times \mathsf{B}_h \times \mathsf{C}_h$ with the norm Y. The motivation for taking B2 in (36a) is mainly $H^2(D)$-conformity required for the regularity (12). It also interacts well with the choice D0 in (36b) because D0 simultaneously approximates functions in B2 and their second derivatives well, which plays a role in (42) and leads to the dimension formula (38) below. More generally, instead of the P2–D1 combination in (36a)–(36b) one can take $\mathrm{P}(p+1)$–$\mathrm{D}(p)$ for any degree $p \geq 0$, so that (42) below corresponds to a so-called continuous Galerkin time-stepping scheme. The choice of the spatial component in (36c) allows integration by parts in space in an expression like $\langle \mathbf{m}_h, -\nabla\phi_h \rangle$ and approximates $\nabla\phi_h$ well. Finally, the space in (36d) is simply the trace of (36a) at $t = T$, and is sufficiently regular for the proximal step (49) to be well-defined.

We impose the no-flux boundary conditions for the density ρ through (cf. (2) and (23)):

$$\text{homogeneous Neumann spatial boundary conditions on } \mathsf{X}_h; \tag{37a}$$

$$\text{homogeneous Dirichlet boundary conditions on the P1 components of } \mathsf{B}_h. \tag{37b}$$

With these boundary conditions, the number of spatial degrees of freedom of X_h (i.e., $\dim \mathrm{B}2^2$) and those of A_h (i.e., $\dim \mathrm{D}0^2$) coincide, and therefore

$$\dim \mathsf{X}_h = \dim \mathsf{A}_h + \dim \mathsf{C}_h. \tag{38}$$

The operator Λ is approximated by the (injective) operator

$$\Lambda_h : \mathsf{X}_h \to \mathsf{Y}_h, \quad \Lambda_h \phi := (Q_1(\partial_t + \nu^2\Delta)\phi), Q_2 \nabla\phi, -\phi(T)), \tag{39}$$

where Q_1 is the L_2-orthogonal projection onto A_h and Q_2 is the componentwise L_2-orthogonal projection onto B_h. We insert these projections for the update Step C §3.2.3 to make sense. They have no effect in the term $(\Lambda\phi_h - \sigma_h, \lambda_h)_{\mathsf{Y}}$ but they do affect the r-term of the augmented Lagrangian (24).

Our aim is therefore to solve the discrete optimization problem (cf. (15))

$$\boxed{\operatorname*{argmin}_{(\phi_h,\sigma_h)\in \mathsf{X}_h\times\mathsf{Y}_h} \{\mathscr{F}_h(\phi_h) + \mathscr{G}_h(\sigma_h)\} \quad \text{s.t.} \quad \sigma_h = \Lambda_h\phi_h} \tag{40}$$

with $\mathscr{F}_h := \mathscr{F}$ and $\mathscr{G}_h := \mathscr{G}$ from (16) (or some convex approximations thereof). Analogously to (18), the convex dual problem now reads

$$\inf_{\lambda_h'\in\mathsf{Y}_h'} \{\mathscr{F}_h^\star(-\Lambda_h'\lambda_h') + \mathscr{G}_h^\star(\lambda_h')\}, \tag{41}$$

with $\mathscr{F}_h^\star(-\Lambda_h'\lambda_h') =$

$$\sup_{\phi_h\in\mathsf{X}_h} \{-\langle\rho_h, (\partial_t + \nu^2\Delta)\phi_h\rangle - \langle \mathbf{m}_h, \nabla\phi\rangle - \langle\rho_0, \phi_h(0)\rangle + \langle e_h', \phi_h(T)\rangle\}, \tag{42}$$

where, notably, the supremum is taken over discrete functions only. Here, $\lambda_h' = (\rho_h, \mathbf{m}_h, e_h')$. It is possible to integrate by parts in space on the term $\langle \mathbf{m}_h, -\nabla\phi_h\rangle$ owing to the choice of the discrete spaces (36a) and (36c). Therefore, the supremum (42) is finite (and equals zero) if and only if the triple $(\rho_h, \mathbf{m}_h, e_h') \in \mathsf{Y}_h'$ satisfies the following discrete analog of (1a),

$$\langle\rho_h, (-\partial_t - \nu^2\Delta)\phi\rangle = \langle -\operatorname{div}\mathbf{m}_h, \phi\rangle + \langle\rho_0, \phi(0)\rangle - \langle e_h', \phi(T)\rangle \quad \forall\phi\in\mathsf{X}_h. \tag{43}$$

In particular, fixing $\mathbf{m}_h$ and restricting the test functions to $\phi(T) = 0$, the dimension count (38) suggests that ρ_h is well-defined by (43) and approximates the solution ρ of $\partial_t\rho - \nu^2\Delta\rho = -\operatorname{div}\mathbf{m}_h$ with $\rho(0) = \rho_0$ and $\nu^2\nabla\rho\cdot\mathbf{n} = 0$ on ∂D in the space-time ultraweak sense (spatial and temporal derivatives are on the discrete test function; the analysis can be done along the lines of [7, 5]). The initial condition and the no-flux boundary condition are thereby injected naturally. Integration by parts in time shows that admitting nonzero $\phi(T)$ additionally ensures $\langle e_h', \chi\rangle = \langle\rho_h(T), \chi\rangle$ for all $\chi\in\mathsf{C}_h$, that is e_h' is determined as the $L_2(D)$-orthogonal projection of $\rho_h(T)$ onto C_h.

Consider $\phi := b_n \otimes (1\otimes 1)$, where $b_n \in \mathrm{P}2$ is a quadratic (nonzero) bubble on the n-th temporal element that vanishes outside that element. Using this ϕ in (43) yields $\int_J b_n\partial_t \int_D \rho_h = 0$. Since $\partial_t\int_D\rho_h$ is piecewise constant in time, it must equal zero. This implies mass conservation.

At least when $\rho_0 > 0$ and in $\mathrm{D}0^2$, we can take $\rho_h(t) := \rho_0$ in (43) to find a corresponding $\mathbf{m}_h$ and e_h'. This will give a triple $\lambda_h' = (\rho_h, \mathbf{m}_h, e_h')$ that yields a finite value in (41). Consequently,

the discrete problem (41) admits a minimizer (under reasonable conditions on A, Γ, and L, e.g. as in the numerical examples below).

An alternative to ALG2 is to first discretize the MFGs equations (1), thus obtaining a nonlinear system of evolution equations, and then apply a root-finding method such as the Newton iteration. This route was investigated in [2]. In this case, a linearized discrete version of (1) is to be solved in each iteration. Some drawbacks associated with this approach are: the saddle point structure of the linear problem as opposed to the symmetric positive definite problem (45); the required degree of differentiability of the data (e.g. C^2 for $H(x,\cdot)$ in [2]); the iterates may not respect constraints such as non-negativity of the density, unless they are already sufficiently close to the solution [2, p. 202]; it is difficult to specify the required accuracy for the iterates. It seems therefore meaningful to switch to a Newton iteration once a good guess has been constructed with the ALG2 iteration.

It is tempting to try to solve the MFGs system (1) by iterating the KFP–HJB equations. This works in practice [16], but averaging as in [15] may furnish a provably convergent iteration.

3.2. Discretized ALG2. We run ALG2 from Section 2.5 on the discrete augmented Lagrangian obtained from (40):

$$L_r(\phi_h,\sigma_h,\lambda_h) = \mathscr{F}_h(\phi_h) + \mathscr{G}_h(\sigma_h) + (\Lambda_h\phi_h - \sigma_h, \lambda_h)_{\mathsf{Y}} + \tfrac{r}{2}\|\Lambda_h\phi_h - \sigma_h\|_{\mathsf{Y}}^2. \tag{44}$$

Suppose $\mathscr{F}_h$ and $\mathscr{G}_h$ are closed proper convex functions into $(-\infty,\infty]$, the operator Λ_h is injective, and fix $r>0$ (we will use $r=1$). Assume that there exists a Kuhn–Tucker pair $(-\Lambda_h'\lambda_h', \lambda_h') \in \partial\mathscr{F}_h(\phi_h)\times\partial\mathscr{G}_h(\Lambda_h\phi_h)$. Then the algorithm converges to a solution of the discrete optimization problem (40) and the convergence is robust under perturbations in Step A and Step B if those perturbations decay sufficiently fast with the iteration [17, Theorem 8].

3.2.1. Discrete Step A. The iterate $\phi_h^{(k+1)} \in \mathsf{X}_h$ is defined by the linear variational problem

$$(\Lambda_h\phi_h^{(k+1)}, \Lambda_h\tilde\phi)_{\mathsf{Y}} = (\sigma_h^{(k)} - \tfrac{1}{r}\lambda_h^{(k)}, \Lambda_h\tilde\phi)_{\mathsf{Y}} - \tfrac{1}{r}\mathscr{F}_h(\tilde\phi) \quad \forall\tilde\phi\in\mathsf{X}_h. \tag{45}$$

With $\mathscr{A}_h := \Lambda_h'\Lambda_h$, where the adjoint is w.r.t. the Y scalar product, this can be written as

$$\mathscr{A}_h\phi_h^{(k+1)} = b_h^{(k+1)} := \Lambda_h'(\sigma_h^{(k)} - \tfrac{1}{r}\lambda_h^{(k)}) - \tfrac{1}{r}\mathscr{F}_h. \tag{46}$$

3.2.2. Discrete Step B. The prior is

$$\bar\sigma_h^{(k+1)} = (\bar a_h^{(k+1)}, \bar{\mathbf{b}}_h^{(k+1)}, \bar c_h^{(k+1)}) := \Lambda_h\phi_h^{(k+1)} + \tfrac{1}{r}\lambda_h^{(k)}. \tag{47}$$

The minimization problem (34) is a pointwise minimization in space-time. Even if the data are discrete functions, the minimizer need not lie in the discrete spaces. We first perform the minimization on collocation nodes on each space-time element that are together unisolvent for a piecewise polynomial space. Specifically, we use the 2-node Gauss–Legendre quadrature points on each one-dimensional element (hence 4 nodes per spatial element and 8 nodes per space-time element) to characterize a function in the discrete space $\mathsf{Z}_h := \mathrm{D1}\otimes\mathrm{D1}^2$. We write $\mathscr{N}(\mathsf{Z}_h)\subset J\times D$ for those collocation nodes. Then we project the result onto the original discrete spaces. The procedure is thus as follows.

1. For each collocation node $n\in\mathscr{N}(\mathsf{Z}_h)$ let $(a_h(n),\mathbf{b}_h(n))$ denote the solution to the pointwise minimization problem

$$\operatorname*{argmin}_{(a,\mathbf{b})\in\mathbb{R}\times\mathbb{R}^d} \left\{A^\star(a+H(\mathbf{b})) + \tfrac{r}{2}|(\bar a_h^{(k+1)}(n), \bar{\mathbf{b}}_h^{(k+1)}(n)) - (a,\mathbf{b})|^2\right\}. \tag{48}$$

2. Construct the intermediate $(a_h, \mathbf{b}_h) \in \mathsf{Z}_h \times \mathsf{Z}_h^2$ from the values on the collocation nodes.
3. Project (orthogonally in $L_2 \times [L_2]^d$) the intermediate $(a_h, \mathbf{b}_h)$ onto $\mathsf{A}_h \times \mathsf{B}_h$ to obtain the new iterates $(a_h^{(k+1)}, \mathbf{b}_h^{(k+1)})$.

Let I_1 denote the operator that constructs a $\mathrm{D}1^2$ function from its values on the spatial collocation nodes $\mathscr{N}(\mathrm{D}1^2) \subset D$. The minimization problem (35) is approximated by

$$c_h^{(k+1)} := \operatorname*{argmin}_{c \in \mathsf{C}_h} \left\{ \int_D \mathrm{I}_1 \Gamma^\star(c) + \tfrac{r}{2} \|\bar{c}_h^{(k+1)} - c\|_{\mathsf{V}}^2 \right\}. \tag{49}$$

3.2.3. Discrete Step C. This is the update $\lambda^{(k+1)} := \lambda^{(k)} + r(\Lambda_h \phi_h^{(k+1)} - \sigma_h^{(k+1)})$. Since the range of Λ_h is contained in Y_h by the definition (39), we have $\lambda^{(k+1)} \in \mathsf{Y}_h$ whenever $\lambda^{(k)} \in \mathsf{Y}_h$.

4. Preconditioning the Step A of ALG2.

4.1. Basic preconditioner. In Section 2.5.1 we introduced the symmetric operator $\mathscr{C} : \mathsf{X} \to \mathsf{X}'$ as

$$\langle \mathscr{C}\phi, \phi \rangle := \|\partial_t \phi\|_{J\times D}^2 + \|\phi(T)\|_D^2 + \|\nabla \phi\|_{J\times D}^2 + \nu^4 \|\Delta \phi\|_{J\times D}^2, \tag{33}$$

and argued that it is equivalent to the operator $\mathscr{A} := \Lambda'\Lambda$ uniformly in the diffusion coefficient ν. We discretize this operator to obtain $\mathscr{C}_h$, defined by $\mathscr{C}_h \phi_h := (\mathscr{C}\phi_h)|_{\mathsf{X}_h}$, and use it as a preconditioner for the discrete operator $\mathscr{A}_h = \Lambda_h' \Lambda_h$ in (46)–(53). Recall from (39) that Λ_h includes the projections Q_1 and Q_2 onto discrete spaces. This begs the question whether the equivalence $\mathscr{A}_h \sim \mathscr{C}_h$ is still true on X_h uniformly in the relevant parameters (as it would be without the projections). The answer is a *conditional yes* in the sense of [5], as reported in Figure 1. We give only a sketch of the underlying mechanism here. Note that it suffices to check the first equivalence in

$$\mathscr{A}_h \sim \mathscr{A} \sim \mathscr{C} \sim \mathscr{C}_h \quad \text{on} \quad \mathsf{X}_h. \tag{50}$$

Hence, for $\phi \in \mathsf{X}_h$, we write $\langle \mathscr{A}_h \phi, \phi \rangle = T_{1,h} + T_{2,h}$ with $T_{1,h} = \|Q_1(\partial_t + \nu^2 \Delta)\phi\|_{J\times D}^2 + \|\phi(T)\|_{\mathsf{V}}^2$ and $T_{2,h} = \|Q_2 \nabla \phi\|_{J\times D}^2$. Let T_1 and T_2 be those quantities without the projections, so that $\langle \mathscr{A}\phi, \phi \rangle = T_1 + T_2$. Clearly, $T_{i,h} \le T_i$. Define the hyperbolic and parabolic CFL numbers in terms of the resolution of the space-time mesh as $\mathrm{CFL}_{\mathrm{hyp}} := (\text{time scale})/(\text{length scale})$ $\mathrm{CFL}_{\mathrm{par}} := \nu^2(\text{time scale})/(\text{length scale})^2$. Consider two cases:

- Case $\nu \gtrsim 1$. In this case $T_2 \lesssim T_1$ holds, because the third term T_2 of (32) is controlled by the second. Moreover, provided $\mathrm{CFL}_{\mathrm{par}} \lesssim 1$, the arguments of [5, Sec. 3.2.3] give $T_1 \lesssim T_{1,h}$ and therefore (50). Compare Figure 1 (left) with the implicit midpoint rule in [5, Fig. 1.1].
- Case $\nu \ll 1$. In this case $T_1 \sim T_{1,h}$ and the term T_2 is the problematic one. But $T_2 \lesssim T_{1,h} + T_{2,h}$ holds provided $\mathrm{CFL}_{\mathrm{hyp}} \lesssim 1$ by an argument similar to [5, (3.20)]. See Figure 1 (right).

We believe that the CFL conditions are fundamental to the good performance of the overall method, because the "continuous Galerkin" P($p+1$)–D(p) temporal discretization is used in (36a)–(36b) to solve the parabolic evolution equation (1b) for ϕ_h with test functions like ρ_h. Perhaps adapting the unconditionally stable discretization variant from [5, Sec. 3.4] could alleviate at least the parabolic CFL restriction.

The inversion of this preconditioner becomes more tractable by passing to a temporal basis of P2 in (36a) that is orthogonal with respect to both scalar products

$$(\theta, \tilde{\theta})_1 := (\theta', \tilde{\theta}')_J + \theta(T)\tilde{\theta}(T) \quad \text{and} \quad (\theta, \tilde{\theta})_0 := (\theta, \tilde{\theta})_J. \tag{51}$$

Such a basis $\{\theta_\omega\}_\omega$ is obtained by solving the generalized eigenvalue problem

$$\text{Find} \quad (\theta_\omega, \omega) \in \mathrm{P2} \times \mathbb{R} \quad \text{s.t.} \quad (\theta_\omega, \tilde{\theta})_1 = \omega^2 \, (\theta_\omega, \tilde{\theta})_0 \quad \forall \tilde{\theta} \in \mathrm{P2}. \tag{52}$$

Note that the first pairing is positive definite on $H^1(0,T)$ due to the T-term, hence the eigenvalues are indeed positive. In this temporal basis, the first two terms of (33) amount to the multiplication by ω^2. The preconditioner $\mathscr{C}_h$ is therefore block-diagonal and each block is the $\mathrm{B2}^2$ discretization C_h^ω, given by $C_h^\omega v := (B^\omega v)|_{\mathrm{B2}^2}$, of the spatial symmetric positive definite biharmonic operator $C^\omega := \omega^2 - \Delta + \nu^4 \Delta^2$ with the homogeneous Neumann boundary conditions (and parameterized by the temporal frequency $\omega > 0$). Thus, applying the inverse of the discrete space-time preconditioner $\mathscr{C}_h$ amounts to solving a series of *independent* problems of the form $C_h^\omega u = f$, which can be done in parallel. This is especially relevant as the quality of the preconditioner improves with the temporal mesh refinement (see Figure 1).

What makes this block-diagonalization possible is the fact that in (33) the differential or evaluation operators do not act on the temporal and the spatial variable simultaneously. This is *not* the case for the operator $\mathscr{A}$ itself in (32) because of the mixed term $\nabla\phi(0)$. This was our motivation for omitting this term in the definition of the operator (33).

4.2. Multigrid-in-space preconditioner. Instead of solving $C_h^\omega u = f$ one can replace the inverse of C_h^ω by a multigrid cycle (or another approximation such as the incomplete Cholesky factorization). We follow the geometric multigrid procedure of [21, Section 4.1]. Starting with the finest spatial mesh with $2^{L_x} \times 2^{L_y}$ uniform rectangular elements, the mesh hierarchy is defined by isotropic coarsening of the mesh until there is a dimension with at most two elements. As the prolongation operator we use the natural embedding, the restriction operator is its adjoint. For the pre- and post-smoother we use the "scaled mass matrix smoother", defined as the preconditioned Richardson iteration $\mathbf{v} \mapsto \mathbf{v} + (\lambda_{\max}^\omega \mathbf{M})^{-1}(\mathbf{f} - \mathbf{C}^\omega \mathbf{v})$, where $\mathbf{M}$ is the spatial mass matrix, $\mathbf{C}^\omega = \omega^2 \mathbf{M} + \mathbf{A} + \nu^4 \mathbf{B}$ is the discretization of C_h^ω on the current level and for the given temporal frequency ω, and $\lambda_{\max}^\omega$ is the maximal eigenvalue (precomputed numerically) of the generalized eigenvalue problem $\mathbf{C}^\omega \mathbf{v} = \lambda_{\max}^\omega \mathbf{M} \mathbf{v}$. Note that the choice of the basis is irrelevant here. For the computation of $\lambda_{\max}^\omega$, the observation $\lambda_{\max}^\omega = \lambda_{\max}^0 + \omega^2$ is useful.

The contraction factor of this multigrid is robust in the parameters ν and ω, as well as in the mesh width: see Figure 3.

4.3. Modified preconditioner. We obtained the basic preconditioner in Section 4.1 as the restriction of the operator $\mathscr{C}$ designed in §2.5.1 to be equivalent to $\mathscr{A} = \Lambda' \Lambda$. Could one start directly with the discrete operator $\mathscr{A}_h = \Lambda_h' \Lambda_h$? As in §2.5.1 one obtains $\|\Lambda_h \phi\|_Y^2 = \|\partial_t \phi\|_{J\times D}^2 + \nu^4 \|P_1 \Delta \phi\|_{J\times D}^2 + \|P_2 \nabla \phi\|_{J\times D}^2 + \|\phi(T)\|_D^2 + \nu^2 \|\nabla\phi(0)\|_D^2$ for $\phi \in X_h$. We again omit the last term to define a "template" operator $\mathscr{D}_{h,\mathrm{temp}}$. Let $p_1 : \mathrm{P2} \to \mathrm{D1}$ and $p_2 : \mathrm{P2} \to \mathrm{D0}$ denote the "temporal parts" of the projections P_1 and P_2. We switch to the temporal basis defined by the eigenvalue problem (52) with the new scalar products $(\cdot,\cdot)_1' := (p_1 \cdot, p_1 \cdot)_J$ and $(\cdot,\cdot)_0' := (\cdot,\cdot)_1$. In that temporal basis, the operator $\mathscr{D}_{h,\mathrm{temp}}$ would be block-diagonal with spatial blocks $1 + (\omega')^2(\nu^4\Delta^2 - \Delta)$ if p_1 and p_2 mapped into the same space. Let us *define* the preconditioner $\mathscr{D}_h$ as this operator. As shown in Figure 2, the condition number of the preconditioned system is of order one, provided $\mathrm{CFL}_{\mathrm{hyp}} \lesssim 1$. In particular, the condition numbers are robust in ν, requiring fewer temporal refinements. Of course, the spatial blocks can again be replaced by multigrid.

Despite the superior performance of this modified preconditioner on coarse temporal meshes, the variant from Section 4.1 might be more relevant for large scale computation because it is easier to find a *sparse* transformation to a temporal basis where both scalar products (51) are

approximately diagonal (cf. [4]). For example, in a standard B-spline wavelet basis they will be approximately diagonal up to a block whose size is logarithmic in the number of temporal degrees of freedom due to the boundary term in (51).

4.4. Discussion. It was noted in Section 2.5.1 that in Step A, essentially a PDE of the form $-\partial_t^2\phi + \nu^4\Delta^2\phi - \Delta\phi = \text{RHS}$ has to be solved on the space-time cylinder. Similar problems appear in the literature on numerical methods for optimal control of parabolic PDEs, even though the cost functional is usually simpler than (7). Take for example the work [11], where the quadratic tracking functional $\|y-z\|_{J\times D}^2 + \alpha\|u\|_{J\times D}^2$ is minimized subject to the heat equation $\partial_t y - \Delta y = -u$, $y|_{\partial D} = 0$, for given initial value $y(0)$ and desired state z. The first order optimality system includes the equations $(\partial_t + \Delta)p = z - y$, $p(T) = 0$ and $u = p$. In this example, y, p correspond to ρ, ϕ. Testing the equation for p by $(\partial_t + \Delta)\tilde{p}$ and integrating in space-time leads to $\langle\tilde{\Lambda}p, \tilde{\Lambda}\tilde{p}\rangle := \langle(\partial_t + \Delta)p, (\partial_t + \Delta)\tilde{p}\rangle + \alpha^{-1}\langle p, \tilde{p}\rangle = \text{RHS}$. This is the analog of Equation (29)–(30). With this definition, $\|\tilde{\Lambda}p\|^2 = \|\partial_t p\|_{J\times D}^2 + \|\Delta p\|_{J\times D}^2 + \alpha^{-1}\|p\|_{J\times D}^2 + \|\nabla p(0)\|_D^2$. Thus, a preconditioner along the lines of §2.5.1 and §4.1 can be constructed (cf. also [4, Sec. 4.1.3]). Vice versa, preconditioners developed in that context might apply here with some adaptations.

A number of preconditioners for the Newton iteration (see Section 3.1) were proposed in [2]. At this point it is appropriate to recall that of [2, Algorithm C]. It consists in a) eliminating ρ from the linear saddle point problem thus obtaining a linear PDE for ϕ; b) applying a linear solver such as BiCGstab; and c) using multigrid with Gauss–Seidel smoothing and spatial semi-coarsening only. The principal part of the PDE in a) is again $-\partial_t^2\phi + \nu^4\Delta^2\phi$, but it also includes a (symmetric nonnegative) term of the form $-\operatorname{div}(\rho^{(k)}\partial_{\mathbf{pp}}H(\nabla\phi^{(k)})\nabla\phi)$. The authors report favorable results, but the interpretation and a meaningful comparison is difficult because the number of BiCGstab iterations is used as an indirect measure of the condition number (with possibly mesh-dependent norms in the stopping criterion), and the effect of temporal vs. spatial refinement is not investigated separately. The quality of that preconditioner does seem to improve with increasing diffusion ν, at least for moderate values of ν. By contrast,

- for the basic preconditioner from Section 4.1, the interplay of the equivalences (50) and the discrete spaces (36) requires increasing the temporal resolution with increasing diffusion until $\text{CFL}_{\text{par}} \lesssim 1$ (and $\text{CFL}_{\text{hyp}} \lesssim 1$) to guarantee good preconditioning.
- for the modified preconditioner from Section 4.3, the quality is robust in ν, only requiring $\text{CFL}_{\text{hyp}} \lesssim 1$. However, we do not have a solid theoretical justification.

5. Numerical examples.

5.1. Implementation. We give here some details on the implementation.
As the starting values for the ALG2 iteration we use the zero vector.
The solution to (46) is approximated by the preconditioned conjugate gradient method,

$$\phi_h^{(k+1)} := \phi_h^{(k)} + \text{PCG}[\mathscr{A}_h, (b_h^{(k+1)} - \mathscr{A}_h\phi_h^{(k)}), \epsilon_{\text{pcg}}, \text{maxit}_{\text{pcg}}, \mathscr{C}_h^{-1}], \tag{53}$$

where $\mathscr{C}_h^{-1}$ is the approximation of the inverse of the preconditioner as described in Section 4.1 (basic preconditioner) or Section 4.2 (multigrid preconditioner). The PCG iteration is initialized with the zero vector. We use the relative residual tolerance $\epsilon_{\text{pcg}} = 10^{-1}$, meaning that the iteration terminates once the residual in the $\mathscr{C}_h^{-1}$ norm (defined by $\|r\|^2 := \langle\mathscr{C}_h^{-1}r, r\rangle$) is reduced by the factor ϵ_{pcg}. The maximal iteration number $\text{maxit}_{\text{pcg}} = 100$ is never attained in our computations.

The discrete spatial biharmonic problems in the basic preconditioner (§4.1) and on the coarsest level of the multigrid preconditioner (§4.2) are solved in Matlab with "backslash".

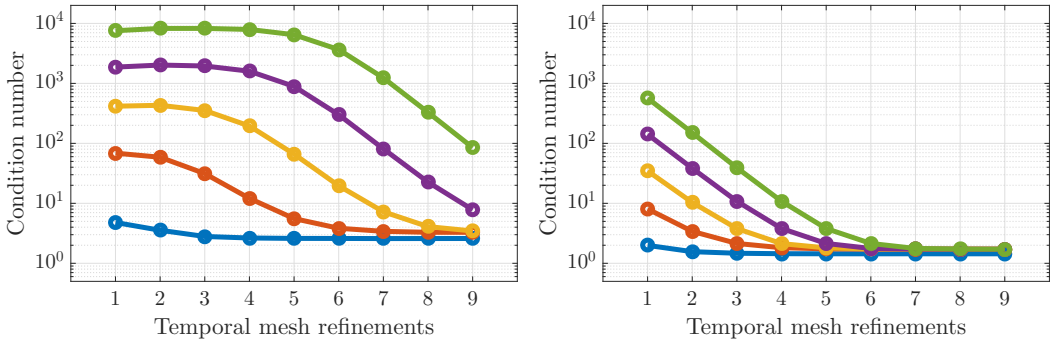

FIG. 1. *The condition number of the discrete operator $\mathscr{A}_h$ preconditioned with the basic preconditioner $\mathscr{C}_h$ from Section 4.1 as a function of the temporal resolution under uniform refinement. The spatial domain is $D = (0,1)^2$, the time horizon is $T = 1$. The lines (bottom-to-top) correspond to $1, \ldots, 5$ uniform spatial refinements.* ***Left:*** *$\nu = 1$.* ***Right:*** *$\nu = 0$.*

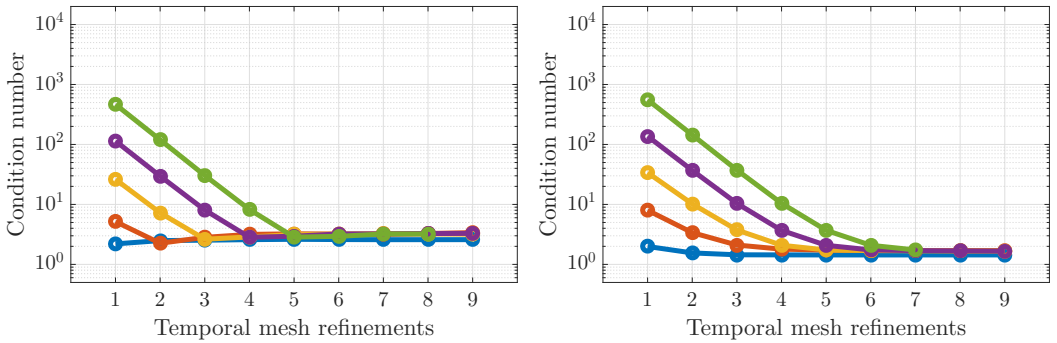

FIG. 2. *As in Figure 1 with the modified preconditioner $\mathscr{D}_h$ from Section 4.3.*

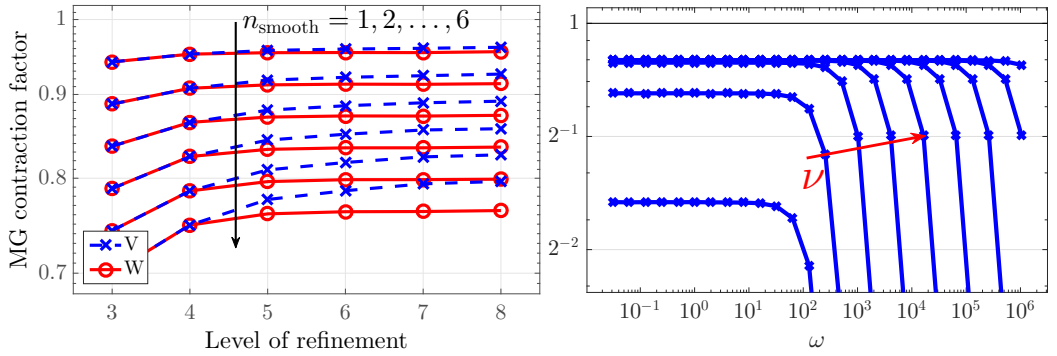

FIG. 3. *Contraction factor $\rho(Id - MG \circ OP)$ of the multigrid from Section 4.2 for $OP := \omega^2 - \Delta + \nu^4\Delta^2$ on a uniform $2^L \times 2^L$ mesh.* ***Left:*** *as a function of the refinement level L and for varying number n_{smooth} of pre- and post-smoothing iterations (V and W cycle), computed as the maximum over $\nu = 2^{-5}, \ldots, 2^3$ and $\omega = 2^{-5}, \ldots, 2^{20}$.* ***Right:*** *as a function of ω for the W cycle with $n_{smooth} = 5$ starting on $L = 8$ and for varying $\nu = 2^{-5}, \ldots, 2^3$.*

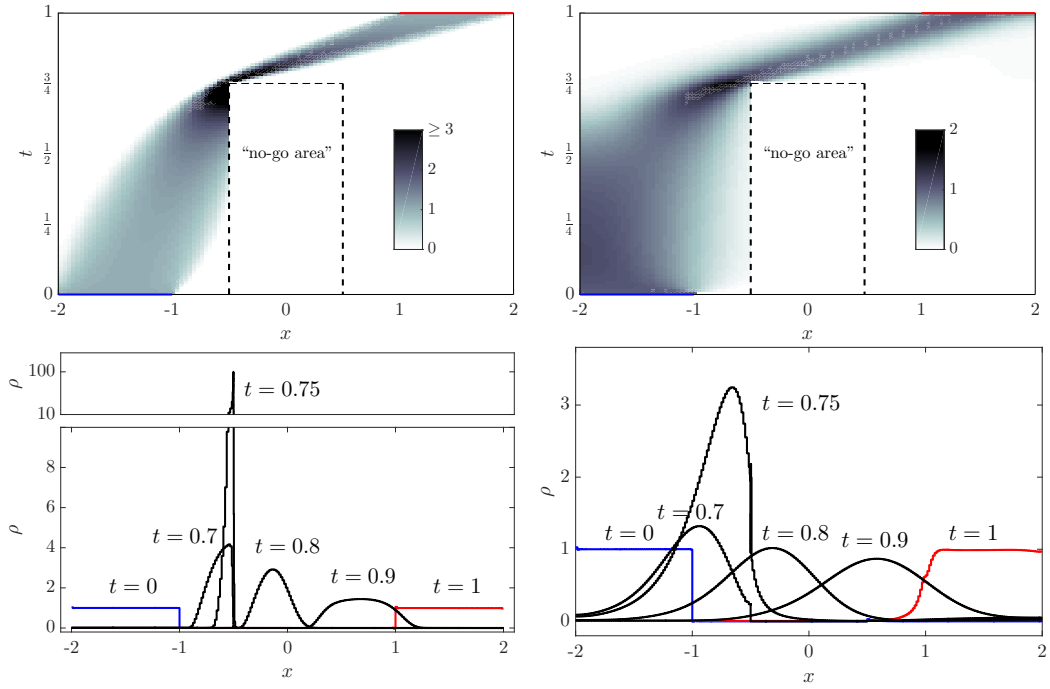

FIG. 4. *The example from Section 5.2. Density ρ in space-time (top) and its temporal snapshots (bottom) with diffusion coefficient $\nu = 10^{-1}$ (left) and $\nu = 1$ (right).*

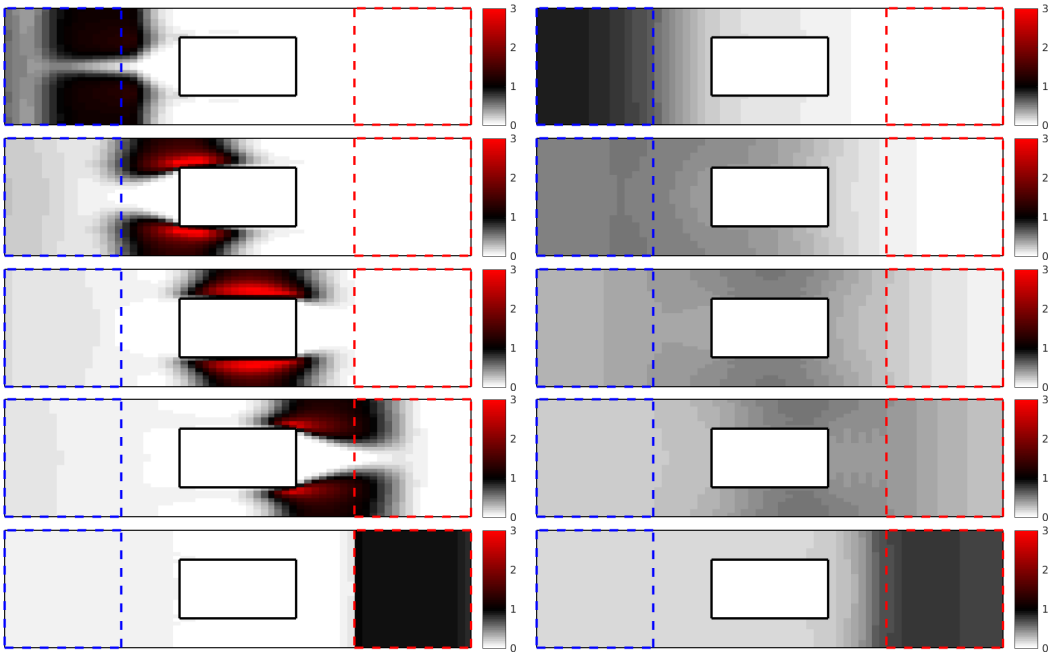

FIG. 5. *The example from Section 5.3. Temporal snapshots $\rho_h(t,\cdot)$ of the computed density at $t = n/13$ for $n = 1, 4, 7, 10, 13$ (top to bottom). Diffusion coefficient $\nu = 10^{-1}$ (left) and $\nu = 1$ (right).*

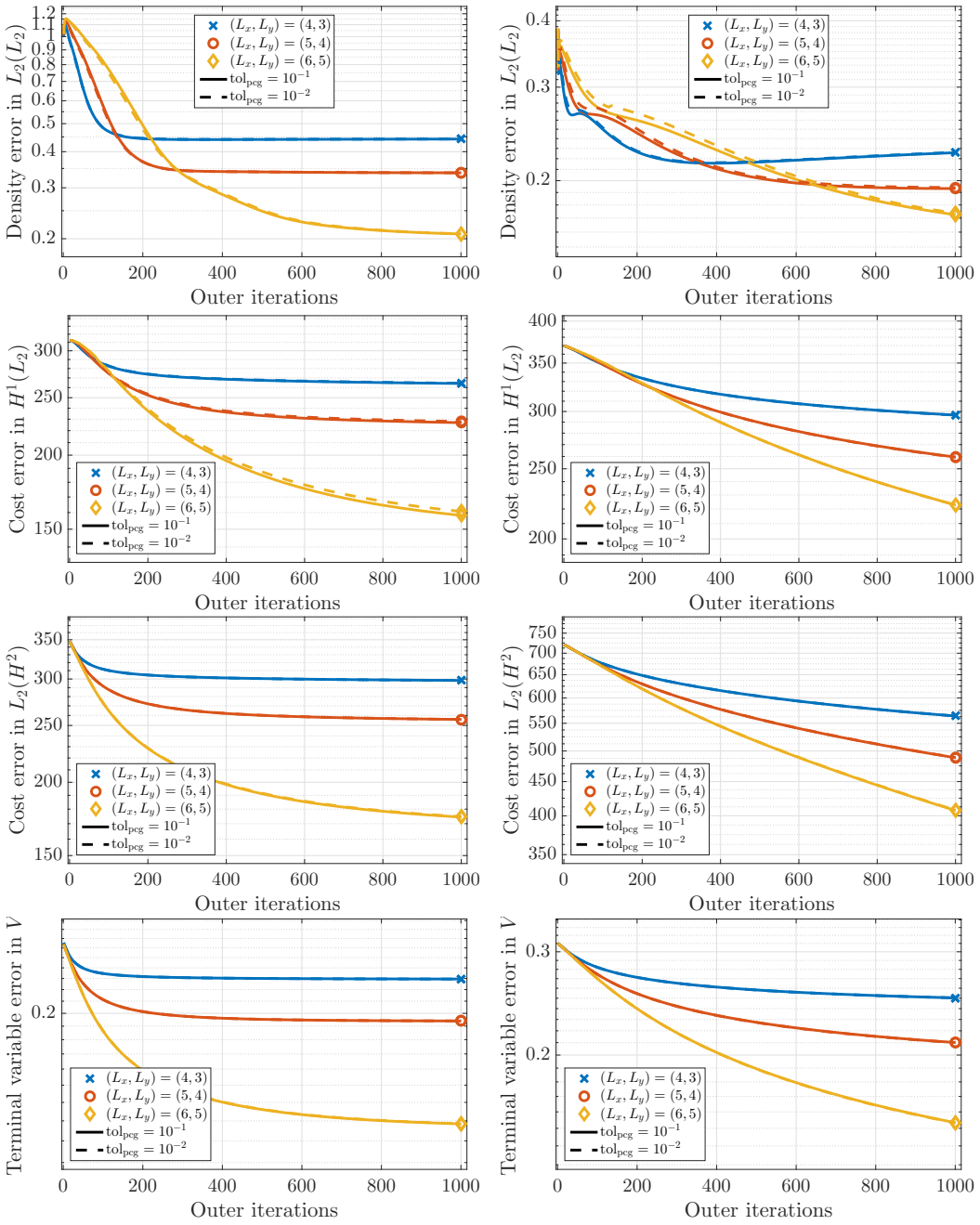

FIG. 6. *Convergence of the discrete ALG2 iteration on the example from Section 5.3 with the basic preconditioner (§4.1) for different spatial resolutions and PCG tolerances in* (53). *Top to bottom: the discrete density* $\rho_h^{(k)}$ *in* $L_2(J; L_2(D))$, *the discrete cost function* $\rho_h^{(k)}$ *in* $H^1(J; L_2(D))$ *and in* $L_2(J; H^2(D))$, *and the terminal variable* $e_h^{(k)}$ *in* V. *Left/Right: diffusion coefficient* $\nu = 10^{-1}$ *and* $\nu = 1$. *The multigrid preconditioner (§4.2) leads to very similar results.*

The eigenvalue $\lambda^{\omega}_{\max}$ in the scaled mass matrix smoother in Section 4.2 is obtained from $\lambda_{\max} = \lambda^0_{\max} + \omega^2$, where $\lambda^0_{\max}$ is approximated using the Matlab `eigs` routine on $\mathbf{M}^{-1}\mathbf{C}^0$ with flag `'LM'`, the options `opt.isreal = 1`, `opt.issym = 0`, and with the default tolerance (the matrices are assembled in the B-spline basis with a modification at the boundary ensuring the homogeneous Neumann boundary condition). Unless specified otherwise, we use the W cycle with $n_{\text{smooth}} = 5$ pre- and post-smoothing steps on each level.

The minimizer of (48) is approximated by a Newton iteration on the derivative of the argument with 20 iterations initializing with the values of the prior. The use of the Newton iteration (rather than the implementation of the proximal operator as in [9, Appendix]) is justified because the second derivative of the functional is still continuous in our examples. This allows to vectorize the Newton iteration in our Matlab implementation, so that Step B takes significantly less time than Step A.

The minimization problem (49) is solved approximately by a Hessian trust-region algorithm implemented in the Matlab routine `fminunc`, and the result is used as the starting value for the same procedure until the relative improvement in the $\|\cdot\|_V$ norm is less than 10^{-10}.

5.2. Example 1. In this example the spatial domain is the interval $D = (-2,2)$ and $T = 1$. The initial density $\rho_0 = \mathbb{1}_{(-2,-1)}$ is the indicator function of $(-2,-1) \subset D$. The terminal cost is $\Gamma(\rho) = 10^3 \times \frac{1}{2}(\rho - \rho_T)^2$, where $\rho_T = \mathbb{1}_{(1,2)}$ is the target terminal state. The congestion cost is $A(\rho) = \frac{1}{2}\rho^2$. Both, Γ and A evaluate to $+\infty$ for $\rho < 0$. The Hamiltonian is $H(t,x,p) = \frac{1}{2}|p|^2 - 10^3 \times \mathbb{1}_{t\le 3/4}\mathbb{1}_{|x|\le 1/2}$, making the movement in the "no-go area" delimited by $|x| \le 1/2$ relatively costly as long as $t \le 3/4$. This leads to the formation of a strong peak of the density $\rho(t,x)$ for $t \nearrow 3/4$ and $x \nearrow -1/2$. In order to resolve this behavior, we geometrically refine the initial $2^8 \times 2^8$ equidistant mesh around $t = 3/4$ by halving the two temporal intervals adjacent to $t = 3/4$ ten times; the same is done for $t = 0$ and for the spatial mesh around $x = -1/2$. The resulting mesh has $(2^8 + 30) \times (2^8 + 20) = 78'936$ space-time elements. We use the basic preconditioner from Section 4.1 (since the spatial domain here is one-dimensional and the mesh is nonuniform). We perform $1'000$ ALG2 iterations. The diffusion coefficient is first set to $\nu = 10^{-1}$. The resulting density ρ behaves as expected: the initial density ρ_0 and the targeted terminal density ρ_T are well-captured, and the "no-go area" $\{t \le 3/4\} \times \{|x| \le 1/2\}$ remains almost mass-free. Now we change the diffusion to $\nu = 1$. The density is now smoothed out in space, in particular eliminating the peak at $t = 3/4$.

One interpretation of this example is as a model for the formation of passenger queue ahead of boarding. The waiting lounge is represented by the interval $(-2,-1)$, the airplane by the interval $(1,2)$, the gate is located at $x = -1/2$, and the gate opening time is $t = 3/4$. In the functional (7), the cost Γ then represents the strong desire to board the plane before takeoff, the A term models the discomfort during queuing, and the L term models the stress of rushing.

5.3. Example 2. Here $D = (-2,2) \times (-1/2, 1/2)$. The initial density is $\rho_0 = \mathbb{1}_{x\le -1}$ in the left part of the domain. The terminal cost is $\Gamma(\rho) := 10^3 \times \frac{1}{2}(\rho - \rho_T)^2$, where $\rho_T := \mathbb{1}_{x\ge 1}$ is the target terminal state in the right part of the domain. The congestion cost is $A(\rho) = \frac{1}{2}\rho^2$. The Hamiltonian is $H(t,x,p) = \frac{1}{2}|p|^2 - 10^3 \times \mathbb{1}_{\square}$, restricting the movement in the rectangular area $\square := \{|x| \le 1/2\} \times \{|y| \le 1/4\}$. The computational mesh has $2^5 \times (2^6 \times 2^5)$ space-time elements. We use the basic preconditioner from Section 4.1. We perform $1'000$ ALG2 iterations. Temporal snapshots of the density for $\nu = 10^{-1}$ and $\nu = 1$ are shown in Figure 5.

5.4. Empirical convergence study. We empirically investigate the convergence of the ALG2 on the example from Section 5.3. As the reference solution we take the discrete solution

computed on the finer mesh with $2^6 \times (2^7 \times 2^6)$ space-time elements, 5′000 ALG2 iterations and the multigrid preconditioner (§4.2). Figure 6 shows the error of the discrete density $\rho_h^{(k)}$ and the discrete cost $\phi_h^{(k)}$ as the ALG2 iteration progresses, varying the spatial resolution (keeping the temporal resolution at 2^5 elements), the PCG tolerance ϵ_{pcg} in (53), and the diffusion coefficient ν. The preconditioner used is the basic preconditioner (§4.1) but the multigrid preconditioner produces (§4.2) essentially the same results. The convergence rate is clearly mesh dependent, indicating nonuniform convexity properties (with respect to the discretization) of the functional to be minimized, see the discussion in Section 2.4. Moreover, for the larger diffusion coefficient $\nu = 1$ we observe a somewhat non-monotonic convergence and 1′000 ALG2 iterations seem not enough; this indicates that even 5′000 ALG2 iterations for the reference solution might be insufficient. The average number of PCG iterations is reported in Table 1.

The application of the preconditioner consumes the bulk of the time. The basic preconditioner (with backslash for the solution of the bi-Laplace equations) takes around 7.5s on the $2^5 \times (2^6 \times 2^5)$ mesh and around 375s on the $2^5 \times (2^7 \times 2^6)$ mesh; the multigrid preconditioner 0.9s and 8s, respectively.

$\nu = 10^{-1}$	Basic, ϵ_{pcg}:		MG, ϵ_{pcg}:	
(L_x, L_y)	10^{-1}	10^{-2}	10^{-1}	10^{-2}
$(4,3)$	3.0	5.0	3.0	5.4
$(5,4)$	3.0	5.3	3.0	6.9
$(6,5)$	3.0	6.7	3.0	7.0

$\nu = 1$	Basic, ϵ_{pcg}:		MG, ϵ_{pcg}:	
(L_x, L_y)	10^{-1}	10^{-2}	10^{-1}	10^{-2}
$(4,3)$	3.0	4.0	4.6	9.8
$(5,4)$	3.0	4.7	3.7	9.7
$(6,5)$	3.0	4.7	3.0	7.2

TABLE 1

Average number of PCG iterations over the first 500 ALG2 steps in the convergence study in Section 5.4. ***Left****:* $\nu = 10^{-1}$. ***Right****:* $\nu = 1$. *For* $\nu = 1$ *with the MG preconditioner, the number of PCG iterations tends to increase in the course of the ALG2 iteration.*

5.5. Intermediate instantaneous costs. Instead of (7) we now minimize

$$J_N(\rho, \mathbf{v}) := \int_{J\times D} \{L(\mathbf{v})\rho + A(\rho)\} + \sum_{i=1}^{N} \int_D \Gamma_i(\rho(\tau_i)) \tag{54}$$

subject to the same transport equation $\mathrm{KFP}[\rho, \mathbf{v}] = 0$. Here, $\{0 =: \tau_0 < \tau_1 < \ldots < \tau_N = T\}$ are temporal nodes where some information on the density ρ is available. Take, say, $\Gamma_i(x, \rho(\tau_i)) := \frac{1}{2}(\rho_{\tau_i}(x) - \rho(\tau_i, x))^2$ for some given spatial densities ρ_{τ_i}. Then the functional (54) selects an evolution of ρ which is propelled by a "lenient" optimal transport starting from ρ_0 along the ρ_{τ_i} and subject to congestion and diffusion effects. The resulting density ρ is not the same as would be obtained by optimizing (7) on the successive temporal intervals $[\tau_{i-1}, \tau_i]$ because more concession in meeting ρ_{τ_i} is potentially made with (54).

To construct a numerical method, the functional framework is modified as follows. We assume $\phi \in \bigtimes_{i=1}^{N} \{\phi|_{[\tau_{i-1},\tau_i]} : \phi \in \mathsf{X}\}$ and $\lambda, \sigma \in L_2 \times [L_2]^d \times \mathsf{V}^N$. The operator Λ is now

$$\Lambda\phi := ((\partial_t + \nu^2\Delta)\phi, \nabla\phi, [\phi]_{\tau_1}, \ldots [\phi]_{\tau_N}), \tag{55}$$

where $[\phi]_t := \phi(t+) - \phi(t-)$ denotes the jump across t with the convention $\phi(T+) := 0$, and the temporal derivative is understood to act on each temporal window $[\tau_{i-1}, \tau_i]$ separately. While $\mathscr{F}$ in (16) remains the same, $\mathscr{G}$ becomes $\mathscr{G}(a, b, c_1, \ldots, c_N) := \int_{J\times D} A^\star(a + H(b)) + \sum_{i=1}^{N} \int_D \Gamma^\star(c_i)$. The discretization is a simple adaptation of (36a)–(36d), allowing temporal

discontinuities in X_h at each τ_i and setting $\mathsf{Y}_h := \mathsf{A}_h \times \mathsf{B}_h \times \mathsf{C}_h^N$. The discrete operator Λ_h includes the projections as in (39). Motivated by (55), the preconditioner for $\Lambda_h'\Lambda_h$ is constructed as in Section 4.1 with the additional terms $\sum_{i=1}^{N-1} [\theta]_{\tau_i} [\tilde{\theta}]_{\tau_i}$ in the definition of $(\theta, \tilde{\theta})_1$ in (51).

As in the example from Section 5.2 we take $D = (-2, 2)$ and the initial density $\rho_0 = \mathbb{1}_{(-2,-1)}$. We set $T = 2$, $\nu = 1$, $A(\rho) = \frac{1}{2}\rho^2$ and $H(t,x,p) = \frac{1}{2}|p|^2$. We define the costs $\Gamma_i(\rho) = 10 \times \frac{1}{2}(\rho - \rho_i)^2$ with $\rho_1 = \mathbb{1}_{(1,2)}$ and $\rho_2 = \rho_0$. Thus we expect the density to accumulate in the interval $(1, 2)$ at time $t = 1$ and then go back to $(-2, -1)$ at time $T = 2$. We compute on a $2^8 \times 2^8$ space-time mesh with the additional temporal tenfold geometric refinement around $t = 1$. We use the basic preconditioner from Section 4.1 with the given modification and perform 1′000 ALG2 iterations. The algorithm appears to be very sensitive to the error incurred by the PCG iteration (53). We start therefore with the initial tolerance $\epsilon_{\text{pcg}} = 10^{-10}$ and increase it by a factor of ten every 10 iterations, leaving it at $\epsilon_{\text{pcg}} = 10^{-2}$ after 80 iterations. The computed density and cost function are shown in Figure 7. The discontinuity of the cost function across $t = 1$ is clearly visible.

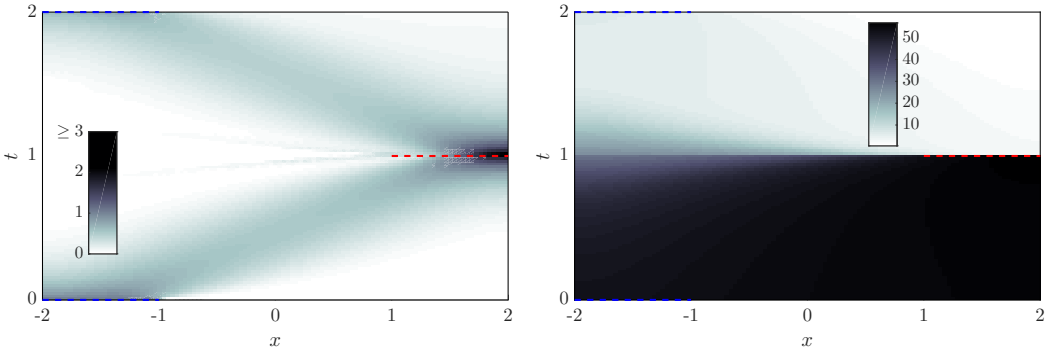

FIG. 7. *Density (left) and cost (right) for the example from Section 5.5.*

Acknowledgment. Thanks to Y. Achdou and J.-D. Benamou for technical advice, and the referees for insightful comments. Support: French ANR-12-MONU-0013 and Swiss NSF #164616.

[1] Yves Achdou and Italo Capuzzo-Dolcetta. Mean field games: numerical methods. *SIAM J. Numer. Anal.*, 48(3):1136–1162, 2010.

[2] Yves Achdou and Victor Perez. Iterative strategies for solving linearized discrete mean field games systems. *Netw. Heterog. Media*, 7(2):197–217, 2012.

[3] Yves Achdou and Alessio Porretta. Convergence of a Finite Difference Scheme to Weak Solutions of the System of Partial Differential Equations Arising in Mean Field Games. *SIAM J. Numer. Anal.*, 54(1):161–186, 2016.

[4] Roman Andreev. Wavelet-in-time multigrid-in-space preconditioning of parabolic evolution equations. *SIAM J. Sci. Comput.*, 38(1):A216–A242, 2016.

[5] Roman Andreev and Julia Schweitzer. Conditional space-time stability of collocation Runge–Kutta for parabolic evolution equations. *Electron. Trans. Numer. Anal.*, 41:62–80, 2014.

[6] Wolfgang Arendt and Ralph Chill. Global existence for quasilinear diffusion equations in isotropic nondivergence form. *Ann. Sc. Norm. Super. Pisa Cl. Sci. (5)*, 9(3):523–539, 2010.

[7] Ivo Babuška and Tadeusz Janik. The h-p version of the finite element method for parabolic equations. II. The h-p version in time. *Numer. Meth. Part. D. E.*, 6:343–369, 1990.

[8] Jean-David Benamou and Yann Brenier. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numer. Math.*, 84(3):375–393, 2000.

[9] Jean-David Benamou and Guillaume Carlier. Augmented Lagrangian methods for transport optimization, mean field games and degenerate elliptic equations. *J Optimiz Theory App*, pages 1–26, 2015.

[10] Jean-David Benamou, Guillaume Carlier, and Filippo Santambrogio. Variational mean field games, 2016. To be published in a special volume on “active particles”.

[11] Alfio Borzì. Multigrid methods for parabolic distributed optimal control problems. *J. Comput. Appl. Math.*, 157(2):365–382, 2003.

[12] Luis M. Briceño-Arias, Dante Kalise, and Francisco J. Silva. Proximal methods for stationary mean field games with local couplings. *arXiv*, 2016. 1608.07701.

[13] Fabio Camilli and Francisco Silva. A semi-discrete approximation for a first order mean field game problem. *Netw. Heterog. Media*, 7(2):263–277, 2012.

[14] Pierre Cardaliaguet, P. Jameson Graber, Alessio Porretta, and Daniela Tonon. Second order mean field games with degenerate diffusion and local coupling. *NoDEA Nonlinear Differential Equations Appl.*, 22(5):1287–1317, 2015.

[15] Pierre Cardaliaguet and Saeed Hadikhanloo. Learning in mean field games: The fictitious play. *ESAIM: COCV*, 23(2):569–591, 2017.

[16] Elisabetta Carlini and Francisco J. Silva. A semi-Lagrangian scheme for a degenerate second order mean field game system. *Discrete and Continuous Dynamical Systems*, 35(9):4269–4292, 2015.

[17] Jonathan Eckstein and Dimitri P. Bertsekas. On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Math. Programming*, 55(3, Ser. A):293–318, 1992.

[18] Ivar Ekeland and Roger Témam. *Convex analysis and variational problems*, volume 28 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1999.

[19] Michel Fortin and Roland Glowinski. *Augmented Lagrangian methods*, volume 15 of *Studies in Mathematics and its Applications*. North-Holland Publishing Co., Amsterdam, 1983.

[20] Pontus Giselsson and Stephen Boyd. Linear convergence and metric selection for Douglas–Rachford splitting and ADMM. *IEEE Transactions on Automatic Control*, 62(2):532–544, 2017.

[21] Wolfgang Hackbusch. *Multigrid methods and applications*, volume 4. Springer-Verlag, Berlin, 1985. Second printing 2003.

[22] Ralf Hiptmair. Operator preconditioning. *Comput. Math. Appl.*, 52(5):699–706, 2006.

[23] Jean-Michel Lasry and Pierre-Louis Lions. Jeux à champ moyen. II. Horizon fini et contrôle optimal. *C. R. Math. Acad. Sci. Paris*, 343(10):679–684, 2006.

[24] Jean-Michel Lasry and Pierre-Louis Lions. Mean field games. *Jpn. J. Math.*, 2(1):229–260, 2007.

[25] Kent-Andre Mardal and Ragnar Winther. Preconditioning discretizations of systems of partial differential equations. *Numer. Linear Algebra Appl.*, 2010.

[26] Nicolas Papadakis, Gabriel Peyré, and Edouard Oudet. Optimal transport with proximal splitting. *SIAM J. Imaging Sci.*, 7(1):212–238, 2014.